

# A Multi-Service Oriented Multiple-Access Scheme for Next-Generation Mobile Networks

Nassar Ksairi, Stefano Tomasin and M  rouane Debbah

Mathematical and Algorithmic Sciences Lab,

France Research Center, Huawei Technologies Co. Ltd., Boulogne-Billancourt, France.

Emails: {nassar.ksairi, stefano.tomasin, merouane.debbah}@huawei.com

**Abstract**—One of the key requirements for fifth-generation (5G) cellular networks is their ability to handle densely connected devices with different quality of service (QoS) requirements. In this article, we present multi-service oriented multiple access (MOMA), an integrated access scheme for massive connections with diverse QoS profiles and/or traffic patterns originating from both handheld devices and machine-to-machine (M2M) transmissions. MOMA is based on a) establishing separate classes of users based on relevant criteria that go beyond the simple handheld/M2M split, b) class dependent hierarchical spreading of the data signal and c) a mix of multiuser and single-user detection schemes at the receiver. Practical implementations of the MOMA principle are provided for base stations (BSs) that are equipped with a large number of antenna elements. Finally, it is shown that such a massive-multiple-input-multiple-output (MIMO) scenario enables the achievement of all the benefits of MOMA even with a simple receiver structure that allows to concentrate the receiver complexity where effectively needed.

## I. INTRODUCTION

In the aim of deploying the Internet of Things (IoT), designing a unified radio access technique for both machine-to-machine (M2M) communications and handheld mobile devices is a challenging problem. One major issue is the difference in terms of traffic patterns and QoS requirements [1] between these two types of communications. Another issue is the large number of IoT devices required to be simultaneously served. Further difficulties arise from the fact that M2M transmissions do not all have the same QoS profile and traffic characteristics [2], [3].

To address some of these challenges, the Third Generation Partnership Project (3GPP) has started to add M2M-type communications support into the radio access subsystem of LTE. Several proposals emerged within this work. Two of them are dubbed LTE for Machine-Type Communications (LTE-M) and Narrow-band LTE-M (NB LTE-M), each introducing a new user equipment (UE) category, the so-called Cat. 1.4MHz for LTE-M and Cat. 200kHz for NB LTE-M [4]. As their respective names indicate, these new UE categories restrict M2M transmissions to a small subband of the available bandwidth that is orthogonal to the broadband users. The same principle is used in Filtered-OFDM [5] with the difference that in Filtered-OFDM, subband-based filtering is applied to enable the use of a different transmission time interval (TTI) and OFDM numerology on the M2M subband. Other solutions consist in providing a separate network for M2M connections. Examples include LoRa<sup>TM</sup> and SIGFOX<sup>TM</sup> [1], which both operate in

the unlicensed frequency bands. While the physical layer of SIGFOX<sup>TM</sup> is based on frequency-division multiple access (FDMA) with ultra narrow band sub-channels, the technology adopted in LoRa<sup>TM</sup> [6] employs a mix of FDMA and of chirp spread spectrum [7]. However, none of the existing solutions is able to meet the following crucial requirements all at once.

**1. Denser IoT deployment:** The existing proposals offer significant improvement over current cellular standards in terms of support for IoT access but there is a need to support much larger numbers of simultaneous M2M transmissions. This goal should be met without sacrificing the QoS of broadband mobile services.

**2. Multi-class users/services:** Not enough attention has been paid to the different QoS and traffic profiles within the class of IoT devices. Indeed, many of these devices *will not all be battery limited sensors and will not only emit small packets of data* [2]. Moreover, there is a need to distinguish from within the services running on handheld devices those that have traffic characteristics and data rate requirements, e.g. short messages, previously considered to be typical of IoT services. Significant gains in resource utilization efficiency are expected from a multiple-access scheme that treats devices with different QoS profiles, whether handheld devices or IoT machines, as belonging to separate classes of users.

**3. Flexibility in resource assignment:** The new multiple-access scheme should be flexible in assigning resources to the different classes and to the different users within each class.

**4. Efficiency in resource utilization:** The new multiple-access scheme should be more efficient in resource utilization than the orthogonal schemes adopted in all the existing proposals.

## II. MULTI-SERVICE ORIENTED MULTIPLE ACCESS

Consider a cell containing a base station (BS) required to serve  $K$  users in the *uplink*<sup>1</sup>. Assume that these users are grouped into  $L \geq 2$  classes defined as follows.

- **One maximum data rate (HD) class:** Here, HD stands for “high data rate”. For this class the objective is to obtain a data rate *as high as possible* for a given number of users. Typically, these users are associated to data hungry applications on handheld devices.

<sup>1</sup>The MOMA principle can be straightforwardly extended to the downlink.

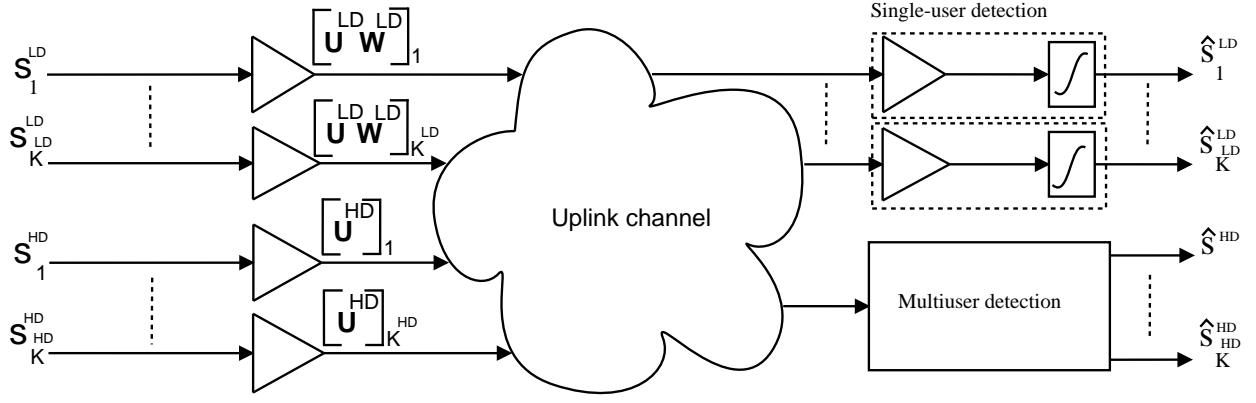


Fig. 1. MOMA transmitters of two classes of users ( $L = 2$ ).

- **$L - 1$  constant data rate (LD) classes:** Here, LD stands for “low data rate”. For these classes, the objective is to accommodate *as many users as possible with granted data rates*  $r_l^{\text{LD}}$  that satisfy

$$r_1^{\text{LD}} > r_2^{\text{LD}} > \dots > r_{L-1}^{\text{LD}}. \quad (1)$$

These users are typically associated with fixed low-rate transmissions from both handheld devices (such as short messages) and from different types of IoT services.

In the sequel, we use  $\mathcal{K}^{\text{HD}} \subset \{1, 2, \dots, K\}$  (resp.  $\mathcal{K}_l^{\text{LD}}$ ) to designate the indexes of the users of the HD (resp. the  $l$ -th LD) class. We also define  $K^{\text{HD}} \stackrel{\text{def}}{=} |\mathcal{K}^{\text{HD}}|$ ,  $K_l^{\text{LD}} \stackrel{\text{def}}{=} |\mathcal{K}_l^{\text{LD}}|$ ,  $\mathcal{K}^{\text{LD}} \stackrel{\text{def}}{=} \bigcup_{l \in \{1 \dots L-1\}} \mathcal{K}_l^{\text{LD}}$ ,  $K^{\text{LD}} \stackrel{\text{def}}{=} |\mathcal{K}^{\text{LD}}|$  and  $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{K}^{\text{HD}} \cup \mathcal{K}^{\text{LD}}$ .

Since what matters for HD users is maximizing throughput, proper scheduling techniques will limit the number of simultaneously-transmitting users, thus it is reasonable to assume that  $K^{\text{HD}}$  will be small. This observation leads us to use multiuser detection techniques for this class of users. On the other hand, LD users will be quite numerous so that multiuser detection would be too complex. Therefore, we propose the use of single-user detection for this class of users. Moreover, to allow for high-performance HD connections, we want to limit interference from LD transmissions on HD received signals, we thus let HD and LD users to be quasi-orthogonal to each other. Lastly, this quasi-orthogonality is implemented in the code domain using a novel *class-dependent hierarchical spreading* scheme.

In the rest of this article we show how such a multiple-access scheme has the desirable property that the available radio resources are used in a *flexible and efficient* manner to allow connecting not only a large number of IoT devices, but also a large number of handheld devices running relatively low data rate services, while guaranteeing the satisfaction of the broadband services with high QoS requirements. This flexibility and efficiency in using the available resources is to be contrasted with the existing access solutions for IoT, such as LoRa<sup>TM</sup>, SIGFOX<sup>TM</sup> and LTE-M, where the resources

reserved for IoT services are under-used and their proportion to the overall resources cannot be dynamically adjusted.

### III. MOMA TRANSMITTER

The MOMA transmitter ( $L = 2$ ) is illustrated in Fig. 1.

#### A. Class Dependent Hierarchical Spreading

Let  $\mathbf{U}$  be an orthogonal-code matrix, e.g. Walsh-Hadamard, whose columns are  $N$ -long spreading codes. In MOMA for  $L$  classes of users, the set of columns of matrix  $\mathbf{U}$  is divided into  $L$  disjoint subsets that form  $L$  matrices, namely  $\mathbf{U}^{\text{HD}}$  and  $\{\mathbf{U}_l^{\text{LD}}\}_{l \in \{1 \dots L-1\}}$  with respective dimensions  $N \times N^{\text{HD}}$  and  $\{N \times N_l^{\text{LD}}\}_{l \in \{1 \dots L-1\}}$ . The  $N^{\text{HD}}$  spreading sequences of  $\mathbf{U}^{\text{HD}}$  will be assigned to  $K^{\text{HD}}$  HD users. Typically,  $K^{\text{HD}} \leq N_l^{\text{LD}}$  so that the HD users could be scheduled in a quasi-orthogonal manner. This inequality could however be violated in the case where the spatial dimension is exploited, as we will see. The  $N_l^{\text{LD}}$  columns of  $\mathbf{U}_l^{\text{LD}}$  will be shared among the users of the  $l$ -th LD class. The scenario of interest for MOMA is when  $K_l^{\text{LD}} > N_l^{\text{LD}}$ , i.e. LD resources are overloaded, and when the following inequalities motivated by (1) are satisfied:

$$\frac{K_{L-1}^{\text{LD}}}{N_{L-1}^{\text{LD}}} > \dots > \frac{K_1^{\text{LD}}}{N_1^{\text{LD}}} > \frac{K^{\text{HD}}}{N^{\text{HD}}}. \quad (2)$$

The transmitted signals of the  $K_l^{\text{LD}}$  users within each LD class  $l$  are formed by means of first linearly combining their data symbols using a rectangular *combining* matrix  $\mathbf{W}_l^{\text{LD}}$  of dimensions  $N_l^{\text{LD}} \times K_l^{\text{LD}}$ , before spreading the resulting symbols using the columns of matrix  $\mathbf{U}_l^{\text{LD}}$ . Applying this class-dependent hierarchical spreading, the final spreading code  $\mathbf{c}_k$  of a LD user  $k$  can be written as

$$\mathbf{c}_k = [\mathbf{U}_l^{\text{LD}} \mathbf{W}_l^{\text{LD}}]_{j_k^{\text{LD}}}, \quad k \in \mathcal{K}_l^{\text{LD}}, l \in \{1 \dots L-1\}, \quad (3)$$

where  $j_k^{\text{LD}} \in \{1, 2, \dots, K_l^{\text{LD}}\}$  is the index of the column of the matrix  $\mathbf{U}_l^{\text{LD}} \mathbf{W}_l^{\text{LD}}$  assigned to user  $k \in \mathcal{K}_l^{\text{LD}}$  and where  $[\mathbf{M}]_{i,j}$  designates the  $i$ -th element of the  $j$ -th column of matrix

M. In principle,  $\mathbf{W}_l^{\text{LD}}$  could be any  $N_l^{\text{LD}} \times K_l^{\text{LD}}$  matrix chosen such that the transmit power constraint is respected:

$$\forall j \in \mathcal{K}_l^{\text{LD}}, \sum_{u=1}^{N_l^{\text{LD}}} \left| [\mathbf{W}_l^{\text{LD}}]_{u,j} \right|^2 = 1. \quad (4)$$

In the following, we assume that the components of  $\mathbf{W}_l^{\text{LD}}$  are chosen as realizations of independent and identically distributed (i.i.d.) random variables that can take the values  $+\frac{1}{\sqrt{N_l^{\text{LD}}}}$  and  $-\frac{1}{\sqrt{N_l^{\text{LD}}}}$  with equal probabilities. In contrast to LD users, no combining is applied to the symbols of the users from the HD class. The spreading code  $\mathbf{c}_k$  used on the signals of a user  $k \in \mathcal{K}^{\text{HD}}$  can thus be written as

$$\mathbf{c}_k = [\mathbf{U}^{\text{HD}}]_{j_k^{\text{HD}}}, \quad k \in \mathcal{K}^{\text{HD}}, \quad (5)$$

where  $[\mathbf{M}]_j$  designates the  $j$ -th column of matrix  $\mathbf{M}$  and where  $j_k^{\text{HD}} \in \{1, 2, \dots, K^{\text{HD}}\}$  is the index of the code assigned to user  $k \in \mathcal{K}^{\text{HD}}$  from among the columns of  $\mathbf{U}^{\text{HD}}$ .

**Remark 1.** *Class dependent hierarchical spreading is distinct from the 2-step spreading (the so-called channelization and scrambling steps [8]) used in third-generation (3G) systems. Indeed, multiplication with  $\mathbf{W}_l^{\text{LD}}$  is intended to precondition signals from a number of users so that they can be spread using a much smaller number of orthogonal codes. This operation is thus completely unrelated, in its conception and in its purpose, to both channelization and scrambling in 3G.*

#### B. MOMA-OFDM

An orthogonal frequency division multiplexing (OFDM) implementation of MOMA (that we designate as MOMA-OFDM) consists in mapping the  $N$  symbols from users' spread signals to  $N \leq N_{\text{FFT}}$  consecutive subcarriers (SCs) in one OFDM symbol, where  $N_{\text{FFT}}$  is the total number of SCs per OFDM symbol. Let  $\mathcal{N}$  be a subset of SCs chosen such that  $|\mathcal{N}| = N$ . Let  $P_k$  designate the average transmit power of user  $k \in \mathcal{K}$ . The signal transmitted by user  $k$  on subcarrier  $n \in \mathcal{N}$  is given by

$$x_{k,n} = \sqrt{P_k} [\mathbf{c}_k]_n s_k, \quad (6)$$

where  $[\mathbf{c}_k]_n$  designates the component of  $\mathbf{c}_k$  mapped to subcarrier  $n$  and where  $s_k$  is the zero-mean unit-variance data symbol transmitted by user  $k$ . Spreading codes  $\mathbf{c}_k$  should be normalized such that  $\mathbb{E}[|x_{k,n}|^2] = \frac{P_k}{N}$ . MOMA-OFDM combines the benefits of both frequency-domain spreading, e.g. the ability to harvest the frequency diversity of the channel and the robustness against carrier frequency shifts, and of OFDM transmission, e.g. robustness against timing errors.

#### C. MOMA with Massive MIMO Base Stations

We now turn our attention to the case where BS is equipped with a large number  $M \gg 1$  of antennas while each user is equipped with a single antenna<sup>2</sup>. This massive MIMO scenario, which is expected to be prevalent in 5G networks,

proves to be particularly advantageous for MOMA, from both the performance and the receiver complexity perspectives. We designate this implementation of MOMA as MIMO-MOMA. Due to the spatial multiplexing capabilities inherent to this scenario, the number of HD users would typically be larger than the number of available HD orthogonal codes, i.e.  $K^{\text{HD}} > N^{\text{HD}}$ . The  $N^{\text{HD}}$  columns of matrix  $\mathbf{U}^{\text{HD}}$  should thus be shared among the  $K^{\text{HD}}$  HD users in such a way that each column is reused by  $\lceil K^{\text{HD}}/N^{\text{HD}} \rceil$  users.

#### IV. MOMA RECEIVER

Denote by  $\mathbf{h}_{k,n}$  the vector of frequency-domain small-scale fading coefficient at subcarrier  $n$  between user  $k$  and the  $M$  antennas of the BS during the current OFDM symbol and assume that the components of  $\mathbf{h}_{k,n}$  are zero-mean i.i.d. random variables and that  $\forall a \in \{1, 2, \dots, M\}, \forall k \in \mathcal{K}, \forall n, m \in \mathcal{N}, \mathbb{E}[[\mathbf{h}_{k,n}]_a^* [\mathbf{h}_{k,m}]_a] = c_{n-m}^h$  where  $\{c_u^h\}_{u \in \mathbb{Z}}$  is a frequency domain autocorrelation sequence. Finally, assume that coefficients  $\mathbf{h}_{k,n}$  can be estimated at the BS by relying on uplink pilot sequences sent by the different user terminals. The vector  $\mathbf{y}_n$  of samples received at the  $M$  BS antennas at subcarrier  $n$  is given by

$$\mathbf{y}_n = \sum_{k \in \mathcal{K}} \sqrt{g_k P_k} \mathbf{h}_{k,n} [\mathbf{c}_k]_n s_k + \mathbf{v}_n, \quad (7)$$

where  $\mathbf{v}_n$  is a  $M \times 1$  vector of i.i.d.  $\mathcal{CN}(0, \sigma^2)$  noise samples and  $g_k$  is the large-scale fading factor. The proposed receiver scheme consists in performing the following operations.

**Spatial demultiplexing** We propose the use of linear receive combining, e.g. according to maximum-ratio (MRC) or minimum-mean-square-error (MMSE) criteria. Combining with coefficients  $\mathbf{d}_{k,n}$  provides the samples  $r_{k,n} \stackrel{\text{def}}{=} \frac{1}{M} \mathbf{d}_{k,n}^H \mathbf{y}_n$ :

$$r_{k,n} = \sqrt{g_k P_k} [\tilde{\mathbf{c}}_k]_n s_k + \sum_{j \neq k} \sqrt{g_j P_j} [\tilde{\mathbf{c}}_j]_n s_j + [\tilde{\mathbf{v}}_k]_n, \quad (8)$$

where we defined for any  $k, j \in \mathcal{K}$  and  $\{n_1, n_2, \dots, n_N\} = \mathcal{N}$

$$\tilde{\mathbf{c}}_j \stackrel{\text{def}}{=} \frac{1}{M} [\mathbf{d}_{k,n_1}^H \mathbf{h}_{j,n_1} [\mathbf{c}_j]_{n_1} \cdots \mathbf{d}_{k,n_N}^H \mathbf{h}_{j,n_N} [\mathbf{c}_j]_{n_N}]^T \quad (9)$$

$$\tilde{\mathbf{v}}_k \stackrel{\text{def}}{=} \frac{1}{M} [\mathbf{d}_{k,n_1}^H \mathbf{v}_{n_1} \cdots \mathbf{d}_{k,n_N}^H \mathbf{v}_{n_N}]^T \quad (10)$$

In the general case where  $\{h_{k,n}\}_{n \in \mathcal{N}}$  are not fully correlated, users' channels are selective in frequency. This implies that the effective spreading codes  $\tilde{\mathbf{c}}_k$  and  $\tilde{\mathbf{c}}_j$  of users  $k$  and  $j$  from two different classes are no longer orthogonal.

**Remark 2** (Channel Hardening). *The effect of receive combining is averaging out small-scale fading over the array, in the sense that the variance of the effective scalar channel gain  $\frac{1}{M} \mathbf{d}_{k,n}^H \mathbf{h}_{k,n}$  decreases with  $M$ . This effect is known as channel hardening and is a consequence of the law of large numbers [10]. the frequency response  $\frac{1}{M} \mathbf{d}_{k,n}^H \mathbf{h}_{k,n}$  of the effective channel is thus asymptotically flat with  $M$  and the above-mentioned loss of orthogonality vanishes.*

**HD Users Detection:** In order to obtain HD data rates that are as high as possible, multiuser detection should be

<sup>2</sup>Extension to the case of multi-antenna user terminals is possible.

used. However, we limit the use of multiuser detection to cases where it is effectively needed. We know from the literature [9] that, under the assumption of uncorrelated antenna channel coefficients, the effective HD spreading gain is  $MN^{\text{HD}}$ . Therefore, we propose the use of MMSE with successive interference cancellation (SIC) *only* in the case where  $K^{\text{HD}} \sim MN^{\text{HD}}$ , as opposed to  $K^{\text{HD}} \ll MN^{\text{HD}}$  for which single-user detection should be sufficient. Define  $\mathbf{A} \stackrel{\text{def}}{=} \text{diag}\{\sqrt{P_k g_k}\}_{k \in \mathcal{K}^{\text{HD}}}$  and the matrix  $\tilde{\mathbf{C}}$  of effective codes as  $\tilde{\mathbf{C}} \stackrel{\text{def}}{=} [\tilde{\mathbf{c}}_{k_1} \cdots \tilde{\mathbf{c}}_{k_{K^{\text{HD}}}}]$ , where  $\{k_1, k_2, \dots, k_{K^{\text{HD}}}\} = \mathcal{K}^{\text{HD}}$ . Now assume that the columns of  $\mathbf{A}$  and  $\tilde{\mathbf{C}}$  are arranged in the descending order with respect to the values  $\sqrt{P_k g_k}$  and define  $\mathbf{T} \stackrel{\text{def}}{=} \tilde{\mathbf{C}}\mathbf{A}$ . The MMSE-SIC receiver starts by recovering the symbol of user  $k_1$  for which the value  $\sqrt{P_{k_1} g_{k_1}}$  is the largest by computing  $r_{k_1}^{\text{HD}} \stackrel{\text{def}}{=} \delta_{k_1}^{\text{H}} \mathbf{r}_{k_1}$  where  $\delta_k \stackrel{\text{def}}{=} (\mathbf{T}\mathbf{T}^{\text{H}} + (\sigma^2 + \sum_{j \in \mathcal{K}^{\text{LD}}} g_j P_j) / M\mathbf{I})^{-1} [\mathbf{T}]_k$  and

$$(K^{\text{HD}} \sim MN^{\text{HD}}) \quad r_{k_i}^{\text{HD}} = \sqrt{g_{k_i} P_{k_i}} \delta_{k_i}^{\text{H}} \tilde{\mathbf{c}}_{k_i} s_{k_i} + \delta_{k_i}^{\text{H}} \tilde{\mathbf{v}}_{k_i} + \sum_{j=i+1}^{K^{\text{HD}}} \sqrt{g_{k_j} P_{k_j}} \delta_{k_i}^{\text{H}} \tilde{\mathbf{c}}_{k_j} s_{k_j} + \sum_{j \in \mathcal{K}^{\text{LD}}} \sqrt{g_j P_j} \delta_{k_i}^{\text{H}} \tilde{\mathbf{c}}_j s_j, \quad (11)$$

for  $k_i = k_1$  and  $\mathbf{r}_{k_1} \stackrel{\text{def}}{=} [r_{k_1, n_1} r_{k_1, n_2} \cdots r_{k_1, n_N}]^{\text{T}}$ . Once  $s_{k_1}$  is decoded correctly, the contribution of  $k_1$  can be removed from  $\{r_{k_2, n}\}_{n \in \mathcal{N}}$  in order to detect the data symbol of user  $k_2$ . The  $N$ -long vector of signal samples after this cancellation is denoted as  $\mathbf{r}_{k_2}$ . This SIC procedure continues till the detection of all the HD data symbols. In the case where  $K^{\text{HD}} \ll M \times N^{\text{HD}}$ , MMSE-SIC is dropped and single-user (SU) detection is instead used by computing  $r_k^{\text{HD, SU}} \stackrel{\text{def}}{=} \mathbf{c}_k^{\text{H}} \mathbf{r}_k$ :

$$(K^{\text{HD}} \ll MN^{\text{HD}}) \quad r_k^{\text{HD, SU}} = \sqrt{g_k P_k} \mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_k s_k + \sum_{j \in \mathcal{K} \setminus \{k\}} \sqrt{g_j P_j} \mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j s_j + \mathbf{c}_k^{\text{H}} \tilde{\mathbf{v}}_k. \quad (12)$$

**LD User Detection:** We propose the use of single-user detection for LD users, so that for any  $k \in \mathcal{K}^{\text{LD}}$ , detection is based on the decision sample  $r_k^{\text{LD}} \stackrel{\text{def}}{=} \mathbf{c}_k^{\text{H}} \mathbf{r}_k$  given by

$$r_k^{\text{LD}} = \sqrt{g_k P_k} \mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_k s_k + \sum_{j \neq k} \sqrt{g_j P_j} \mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j s_j + \mathbf{c}_k^{\text{H}} \tilde{\mathbf{v}}_k. \quad (13)$$

#### A. Detection Signal to Noise Plus Interference Ratio

The signal to noise plus interference ratio (SINR) of any LD user  $k \in \mathcal{K}_l^{\text{LD}}$  can be derived from (13) and is given by

$$\text{SINR}_k^{\text{LD}} \stackrel{\text{def}}{=} \frac{g_k P_k |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_k|^2}{\sum_{j \in \mathcal{K} \setminus \{k\}} g_j P_j |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j|^2 + \sigma_{\text{LD}}^2}, \quad (14)$$

where  $\sigma_{\text{LD}}^2 \stackrel{\text{def}}{=} \frac{1}{M^2} \sum_{n \in \mathcal{N}} |[\mathbf{c}_k]_n|^2 \mathbf{d}_{k,n}^{\text{H}} \mathbf{d}_{k,n} \sigma^2$ . The SINR for any  $k \in \mathcal{K}^{\text{HD}}$  when MMSE-SIC is used follows from (11) as

$$\text{SINR}_{k_i}^{\text{HD}} \stackrel{\text{def}}{=} \frac{g_{k_i} P_{k_i} |\delta_{k_i}^{\text{H}} \tilde{\mathbf{c}}_{k_i}|^2}{\sum_{j=i+1}^{K^{\text{HD}}} g_{k_j} P_{k_j} |\delta_{k_i}^{\text{H}} \tilde{\mathbf{c}}_{k_j}|^2 + \sum_{j \in \mathcal{K}^{\text{LD}}} g_j P_j |\delta_{k_i}^{\text{H}} \tilde{\mathbf{c}}_j|^2 + \sigma_{\text{HD}}^2} \quad (15)$$

where  $\sigma_{\text{HD}}^2 \stackrel{\text{def}}{=} \frac{1}{M^2} \sum_{n \in \mathcal{N}} |[\delta_k]_n|^2 \mathbf{d}_{k,n}^{\text{H}} \mathbf{d}_{k,n} \sigma^2$ . While in the case of single-user detection (12) gives rise to

$$\text{SINR}_k^{\text{HD, SU}} \stackrel{\text{def}}{=} \frac{g_k P_k |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_k|^2}{\sum_{j \in \mathcal{K} \setminus \{k\}} g_j P_j |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j|^2 + \sigma_{\text{HD}}^2}, \quad (16)$$

Finally, define the *instantaneous* bits/s/Hz capacities<sup>3</sup>  $R_k^{\text{LD}} \stackrel{\text{def}}{=} \log(1 + \text{SINR}_k^{\text{LD}})$  and  $R_k^{\text{HD, SU}} \stackrel{\text{def}}{=} \log(1 + \text{SINR}_k^{\text{HD, SU}})$ .

#### B. Asymptotic Analysis

The following theorem states that both inter-class and HD intra-class interference become asymptotically negligible as  $M$  increases even if only single-user detection is employed.

**Theorem 1.** Assume that the components of each  $\mathbf{W}_l^{\text{LD}}$  are realizations of i.i.d. zero-mean random variables that satisfy the condition in (4). Also assume that the empirical distribution of the large-scale fading coefficients  $\{g_k\}_{k \in \mathcal{K}}$  converges as  $K \rightarrow \infty$  to the distribution of a random variable with mean  $\mathbb{E}[g]$ . Finally,  $\forall k \in \mathcal{K}^{\text{LD}}, P_k = P^{\text{LD}}$ . If  $\frac{K^{\text{LD}}}{M} \rightarrow_{M \rightarrow \infty} \alpha$  while  $K^{\text{HD}} = \mathcal{O}_M(1)$  and  $N \ll N_{\text{FFT}}$ , then we have as  $M \rightarrow \infty$

$$(k \in \mathcal{K}^{\text{HD}}) \quad \sum_{j \in \mathcal{K} \setminus \{k\}} g_j P_j |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j|^2 \xrightarrow{p} 0, \quad (17)$$

$$(k \in \mathcal{K}_l^{\text{LD}}) \quad \sum_{j \in \mathcal{K} \setminus \{k\}} g_j P_j |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j|^2 - \frac{c K_l^{\text{LD}}}{N_l^{\text{LD}} M} \xrightarrow{p} 0, \quad (18)$$

$$(k \in \mathcal{K}) \quad \mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_k \xrightarrow{a.s.} 1, \quad (19)$$

where  $c \stackrel{\text{def}}{=} P^{\text{LD}} \mathbb{E}[g]$  and where  $\xrightarrow{p}$  and  $\xrightarrow{a.s.}$  stand for convergence in probability and for almost sure convergence of random variables, respectively.

*Proof.* Assume that  $L = 2$  so that we can drop from now on the use of the index  $l$ . This assumption is only made for the sake of ease of presentation as the following proof arguments apply in the general case of  $L \geq 2$ .

To show that (17) holds, we write its left-hand side as

$$\begin{aligned} \sum_{j \in \mathcal{K} \setminus \{k\}} g_j P_j |\mathbf{c}_k^{\text{H}} \tilde{\mathbf{c}}_j|^2 &= \sum_{j \in \mathcal{K}^{\text{HD}} \setminus \{k\}} g_j P_j \left| \sum_{n \in \mathcal{N}} [\mathbf{c}_k]_n \frac{1}{M} \mathbf{h}_{k,n}^{\text{H}} \mathbf{h}_{j,n} [\mathbf{c}_j]_n \right|^2 + \\ &\quad \sum_{j \in \mathcal{K}^{\text{LD}}} g_j P_j \left| \sum_{n \in \mathcal{N}} [\mathbf{c}_k]_n \frac{1}{M} \mathbf{h}_{k,n}^{\text{H}} \mathbf{h}_{j,n}^{\text{H}} \sum_{u=1}^{N^{\text{LD}}} [\mathbf{U}^{\text{LD}}]_{n,u} [\mathbf{W}^{\text{LD}}]_{u,j} \right|^2 \end{aligned} \quad (20)$$

The first term in (20) converges almost surely (a.s.) to zero as  $M \rightarrow \infty$  due to applying the law of large numbers to the

<sup>3</sup>log denotes the base-2 logarithm.

sum  $\frac{1}{M} \mathbf{h}_{k,n}^H \mathbf{h}_{j,n}$ . As for the second term, it can be rewritten by referring to (3) as

$$\begin{aligned} & \sum_{j \in \mathcal{K}^{\text{LD}}} g_j P_j \left| \sum_{n \in \mathcal{N}} [\mathbf{c}_k]_n \frac{1}{M} \mathbf{h}_{k,n}^H \mathbf{h}_{j,n} \sum_{u=1}^{N^{\text{LD}}} [\mathbf{U}^{\text{LD}}]_{n,u} [\mathbf{W}^{\text{LD}}]_{u,j} \right|^2 \\ &= P^{\text{LD}} \sum_{u,v=1}^{N^{\text{LD}}} \sum_{n,m \in \mathcal{N}} [\mathbf{c}_k]_n [\mathbf{c}_k]_m^* [\mathbf{U}^{\text{LD}}]_{n,u} [\mathbf{U}^{\text{LD}}]_{m,v} \times \\ & \frac{1}{M^2} \sum_{j \in \mathcal{K}^{\text{LD}}} g_j \mathbf{h}_{k,n}^H \mathbf{h}_{j,n} \mathbf{h}_{j,m}^H \mathbf{h}_{k,m} [\mathbf{W}^{\text{LD}}]_{u,j} [\mathbf{W}^{\text{LD}}]_{v,j}^* . \end{aligned} \quad (21)$$

Defining  $\forall n \in \mathcal{N}$ ,  $\xi_{k,j,n} \stackrel{\text{def}}{=} \frac{1}{\sqrt{M}} \mathbf{h}_{k,n}^H \mathbf{h}_{j,n}$  we can write

$$\begin{aligned} & \frac{1}{M^2} \sum_{j \in \mathcal{K}^{\text{LD}}} g_j \mathbf{h}_{k,n}^H \mathbf{h}_{j,n} \mathbf{h}_{j,m}^H \mathbf{h}_{k,m} [\mathbf{W}^{\text{LD}}]_{u,j} [\mathbf{W}^{\text{LD}}]_{v,j}^* = \\ & \frac{K^{\text{LD}}}{M} \frac{1}{K^{\text{LD}}} \sum_{j \in \mathcal{K}^{\text{LD}}} g_j \xi_{k,j,n} \xi_{k,j,m}^* [\mathbf{W}^{\text{LD}}]_{u,j} [\mathbf{W}^{\text{LD}}]_{v,j}^* . \end{aligned} \quad (22)$$

Now, thanks to the assumption that the empirical distribution of  $\{g_k\}_{k \in \mathcal{K}}$  converges as  $K \rightarrow \infty$  to the distribution of a random variable with mean  $\mathbb{E}[g]$  and that the components of  $\mathbf{W}^{\text{LD}}$  are realizations of i.i.d. zero-mean random variables, the arguments of the proof of Proposition 3.3 from [11] can be applied to show, after tedious but straightforward steps, that for each value of  $(n, m, u, v) \in \mathcal{N}^2 \times \{1, 2, \dots, N^{\text{LD}}\}^2$

$$\begin{aligned} & \frac{1}{K^{\text{LD}}} \sum_{j \in \mathcal{K}^{\text{LD}}} g_j \xi_{k,j,n} \xi_{k,j,m}^* [\mathbf{W}^{\text{LD}}]_{u,j} [\mathbf{W}^{\text{LD}}]_{v,j}^* \xrightarrow{P} \\ & \mathbb{E}[g] \mathbb{E}[\xi_{k,j,n} \xi_{k,j,m}^*] \mathbb{E}[[\mathbf{W}^{\text{LD}}]_{u,j} [\mathbf{W}^{\text{LD}}]_{v,j}^*] = \quad (23) \\ & \frac{1}{N^{\text{LD}}} \mathbb{E}[g] |c_{n-m}^h|^2 \delta_{u,v} , \end{aligned}$$

where  $\{c_u^h\}_{u \in \mathbb{Z}}$  is the frequency domain autocorrelation sequence of users' channels and where  $\delta_{u,v} = 1$  if  $u = v$  and  $\delta_{u,v} = 0$  otherwise. Note that  $\forall n, m \in \mathcal{N}$ ,  $|c_{n-m}^h|^2 \approx 1$  thanks to the assumption that  $N \ll N_{\text{FFT}}$ . Plugging  $|c_{n-m}^h|^2 = 1$  and (23) into (21), we get

$$\begin{aligned} & \sum_{j \in \mathcal{K}^{\text{LD}}} g_j P_j \left| \sum_{n \in \mathcal{N}} [\mathbf{c}_k]_n \frac{1}{M} \mathbf{h}_{k,n}^H \mathbf{h}_{j,n} \sum_{u=1}^{N^{\text{LD}}} [\mathbf{U}^{\text{LD}}]_{n,u} [\mathbf{W}^{\text{LD}}]_{u,j} \right|^2 \\ & \xrightarrow{P} \frac{\alpha P^{\text{LD}} \mathbb{E}[g]}{N^{\text{LD}}} \sum_{u=1}^{N^{\text{LD}}} \sum_{n,m \in \mathcal{N}} [\mathbf{c}_k]_n [\mathbf{c}_k]_m^* [\mathbf{U}^{\text{LD}}]_{n,u} [\mathbf{U}^{\text{LD}}]_{m,u} \\ &= \frac{\alpha P^{\text{LD}} \mathbb{E}[g]}{N^{\text{LD}}} \sum_{u=1}^{N^{\text{LD}}} |\mathbf{c}_k^H [\mathbf{U}^{\text{LD}}]_u|^2 \\ &= 0 , \end{aligned} \quad (24)$$

where the last equality follows from the fact that the HD spreading code  $\mathbf{c}_k$  is orthogonal by construction to  $[\mathbf{U}^{\text{LD}}]_u$  for any  $u \in \{1, 2, \dots, N^{\text{LD}}\}$ . Using similar arguments, one can show that (18) holds true. Finally, (19) holds due to (3)

and (5) and to the law of large numbers applied to the sum  $\frac{1}{M} \mathbf{h}_{k,n}^H \mathbf{h}_{k,n}$ . This completes the proof of Theorem 1.  $\square$

Applying Theorem 1 along with the continuous-mapping theorem to (16) reveals that  $R_k^{\text{HD,SU}}$  is increasing with  $M$  in an unbounded manner. As for LD users, the following result can be used to properly tune  $P^{\text{LD}}$ ,  $K_l^{\text{LD}}$  and  $N_l^{\text{LD}}$  so that  $\forall l \in \{1 \dots L-1\}$  the target  $r_l^{\text{LD}}$  is achieved.

**Corollary 1.** *Under the assumptions made in Theorem 1,  $R_k^{\text{LD}} - \log(1 + \text{SINR}_k^{\text{LD},\infty}) \xrightarrow{P} 0$ , where*

$$\forall k \in \mathcal{K}_l^{\text{LD}}, \quad \text{SINR}_k^{\text{LD},\infty} \stackrel{\text{def}}{=} \frac{g_k P^{\text{LD}}}{\frac{c K_l^{\text{LD}}}{N_l^{\text{LD}} M} + \frac{\sigma^2}{M}} . \quad (25)$$

## V. NUMERICAL RESULTS

Simulations results were obtained assuming users' distances to the BS are randomly chosen from the interval  $[25, 100]$  m and that the associated pathloss coefficients  $g_k$  are computed using the COST-231 Hata model [13] with a carrier frequency  $f_0 = 900$  MHz. Users' transmit power is equal to 23 dBm while the noise power spectral density is equal to  $N_0 = -174$  dBm/Hz. Two channel models are considered, namely the Extended Pedestrian A (EPA) and the Extended Vehicular A (EVA) models [12]. Channels generated using the EPA model have smaller delay spreads, and hence frequency responses that are less selective, than the EVA model. For the underlying OFDM system we assume a total number  $N_{\text{FFT}} = 1,024$  of subcarriers out of which, as in LTE, only  $N_{\text{SC}} = 600$  SCs are used for data transmission. Furthermore, we consider a 2-class MOMA-OFDM, i.e.  $L = 2$ , that is implemented on the basis of  $N = 32$  subcarriers using a  $32 \times 32$  Walsh-Hadamard matrix, i.e. 19 instances of MOMA-OFDM are needed to cover the  $N_{\text{SC}} = 600$  available SCs. Out of the  $N$  orthogonal codes,  $N^{\text{HD}} = \frac{72}{8N}$  are reserved for HD users and  $N^{\text{LD}} = \frac{1}{8N}$  for LD users. This partition was chosen for the purposes of fair comparison with LTE-M. Indeed, in the latter system 72 SCs are reserved for M2M transmissions, corresponding to  $\frac{72}{600} \approx \frac{1}{8}$  of the available  $N_{\text{SC}} = 600$  SCs. Finally,  $K^{\text{HD}} = 8N^{\text{HD}}$  corresponding to a spatial multiplexing gain of 8. All the following results have been obtained by averaging over 100 realizations of users' random positions in the cell.

Fig. 2 shows the average number of LD users that can be simultaneously served using MIMO-MOMA as function of the LD data rate requirement  $r^{\text{LD}}$  when compared to both LoRa and a narrow-band cellular IoT system implemented using LTE-M parameters. From the figure we can notice the significant advantage of using MOMA as opposed to narrow-band access schemes in terms of IoT network capacity. Indeed, the resources reserved in the latter systems for M2M communications turn out to be under-used when compared to MOMA. Also note that the performance gap between a narrow-band cellular IoT system and MOMA is the largest on the range of low to moderate LD target data rates. As for the range of high target data rates (on which LoRa slightly outperforms the 2-class implementation of MOMA), **introducing a third class**

for high-rate M2M transmissions can in principle cancel this performance gap. It is worth mentioning that the maximum

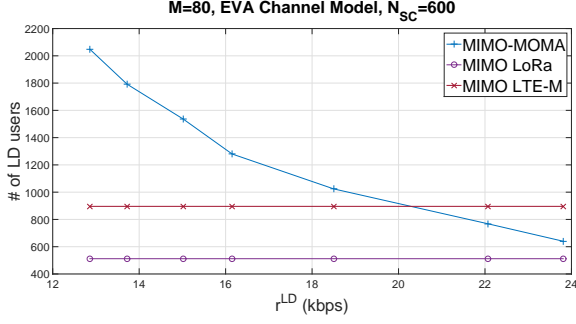


Fig. 2. Number of served LD users within one OFDM symbol vs. the LD data rate requirement.  $M = 80$  with spatial multiplexing.

number of simultaneous M2M transmissions in LoRa was computed assuming both a spatial multiplexing gain of 8 (for fair comparison with our MIMO-MOMA setting) and the availability of 16 125kHz-channels, each of which can support up to 7 concurrent transmissions thanks to the multiplexing capabilities of chirp spread spectrum [6]<sup>4</sup>.

Finally, Fig. 3 shows that MIMO-MOMA achieves HD data rates that are very close to the maximum value achievable with perfect orthogonality. For instance, the HD data rate achieved by MOMA on EPA channels (resp. on EVA channels) with single-user detection stays within 99% (resp. within 86%) of the perfect-orthogonality upper bound. Note that this performance is achieved by MOMA while serving a number of concurrent LD transmissions as large as 4 times the number that can be supported with narrow-band cellular IoT solutions.

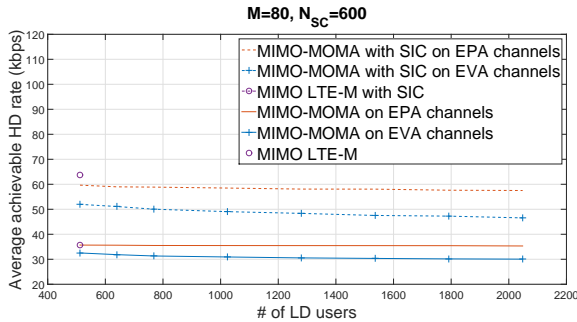


Fig. 3. Achievable rate of HD users vs. the number of served LD users within one OFDM symbol.  $M = 80$  with spatial multiplexing.

## VI. CONCLUSION

In this article, we presented a novel multiple access scheme (MOMA) for next-generation cellular networks that is fully compatible with massive MIMO. This scheme is based on assigning, in a flexible and dynamic manner, different resources

and different degrees of resource overloading to different classes of users, each representing a different data rate requirement and/or a different service type. Moreover, transmissions from different classes in MOMA are quasi-orthogonal. This way, the use of non-orthogonal access for the lower-rate classes would only slightly affect the broadband users, dropping the need for wasteful guard bands, for subband-based filtering or for complex receiver schemes.

## REFERENCES

- [1] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, *Long-Range Communications in Unlicensed Bands: the Rising Stars in the IoT and Smart City Scenarios*. Available: <http://arXiv:1510.00620v1>, Oct. 2015.
- [2] The 3rd Generation Partnership Project (3GPP), *3GPP Technical Specifications 22.368, V.13.0.0, Service Requirements for Machine-Type Communications (MTC): Stage 1*. Available: <http://www.3gpp.org/>, Dec. 2014.
- [3] NGMN, *NGMN 5G White Paper*. Available: <http://www.ngmn.org/>, March 2015.
- [4] Nokia Networks, *LTE-M - Optimizing LTE for the Internet of Things*. Available: <http://networks.nokia.com/>, 2014.
- [5] J. Abdoli, M. Jia, and J. Ma, *Filtered OFDM: A New Waveform for Future Wireless Systems*, in *SPAWC*, Stockholm, July 2015.
- [6] LoRa<sup>TM</sup> Alliance, “LoRa<sup>TM</sup> Specifications V1.0,” Tech. Rep., May 2015.
- [7] G. Naddafzadeh-Shirazi, L. Lampe, G. Vos, and S. Bennett, “Coverage Enhancement Techniques for Machine-to-Machine Communications over LTE,” *IEEE Communications Magazine*, vol. 53, no. 7, July 2015.
- [8] J. Korhonen, *Introduction to 3G Mobile Communications*, 2nd ed. Artech House, Norwood, MA, USA, 2003.
- [9] S. V. Hanly and D. Tse, “Resource Pooling and Effective Bandwidths in CDMA Networks with Multiuser Receivers and Spatial Diversity,” *IEEE Trans. Inf. Theory*, vol. 47, no. 4, May 2001, pp. 1328-1351.
- [10] E. Björnson, E. G. Larsson, and T. L. Marzetta, *Massive MIMO: Ten Myths and One Critical Question*. Available: <http://arXiv:1503.06854v2>, Aug. 2015.
- [11] D. Tse and S. V. Hanly, “Linear Multiuser receivers: Effective Interference, Effective Bandwidth and User Capacity,” *IEEE Trans. Inf. Theory*, vol. 45, no. 2, May 1999, pp. 641-657.
- [12] The 3rd Generation Partnership Project (3GPP), *Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception*. Available: <http://www.3gpp.org/>, Sept. 2015.
- [13] V.S. Abhayawardhana, I.J. Wassell, D. Crosby, M.P. Sellars, and M.G. Brown, *Comparison of Empirical Propagation Path Loss Models for Fixed Wireless Access Systems*, in *VTC Spring*, Stockholm, May 2005, pp. 73-77.

<sup>4</sup>In LoRa, only concurrent transmissions using different spreading factors, and hence having different data rates on the range 0.3~50 kbps, are orthogonal. This explains why the maximum number of simultaneous transmissions in LoRa is plotted as a constant function with respect to the target data rate.